The Journal of Open Source Software

Polyatomic Complexes: A Software Framework for Topologically Accurate Representations of Molecules

Rahul Khorana [1]

1 Department of Computing, Imperial College London, London, United Kingdom

DOI: 10.21105/joss.08828

Software

- Review
- Repository
- Archive

Editor: Rocco Meli

Reviewers:

- @JacksonBurns
- @ritesh001

Submitted: 17 April 2025
Published: 21 October 2025

Summary

Developing robust representations of chemical structures that enable models to learn topological inductive biases is challenging. Polyatomic complexes are a novel learning representation for atomistic systems that addresses this challenge. The representation satisfies numerous structural, geometric, efficiency, and generalizability constraints.

Statement of Need

Current molecular representations, such as SMILES, SELFIES, or graph-based fingerprints, are limited in their ability to accurately reflect electronic structure and higher-order topological features. Additionally, existing representations in computational chemistry fail to satisfy some non-empty subset of accuracy, generalizability, and efficiency constraints (Khorana et al., 2024). Polyatomic Complexes fill this gap by providing a physics-informed and topologically accurate representation that is general across chemical domains, modular, and compatible with standard ML workflows.

Fundamentally, chemical representations should satisfy the following criteria (Langer et al., 2022).

1. Invariances: Representations should be invariant under changes in atom indexing and those fundamental to physics. These invariances are rotation, reflection, and translations (Langer et al., 2022).
2. Uniqueness: Essentially, two systems differing in properties should be mapped to different representations. Systems with equal representations that differ in property induce errors. Uniqueness is necessary and sufficient for reconstruction, up to invariant transformations, of an atomistic system from its representation (Langer et al., 2022).
3. Continuity and Differentiability: Representations of atomistic systems should be continuous and differentiable with respect to atomic coordinates (Langer et al., 2022). Moreover, discontinuities work against regularity assumptions of many machine learning models.
4. Generality: We say a representation of atomistic systems or molecules satisfies generality only if it can encode any atomistic system (Langer et al., 2022).
5. Efficiency: Essentially, representing atomistic systems should be computationally efficient. Ideally, representations are linear in the number of elements in a molecule, $O(S)$, as is the case with molecular graphs (Krenn et al., 2020).
6. Topological Accuracy: Representations are topologically accurate if they can correctly represent the geometry of any molecule or atomistic system. Correctness requires representing the shape, bond-angles, dihedrals/torsion, and electronic structure aspects accurately (Khorana et al., 2024).
7. Long-range interactions: The term long-range interactions refers to electrostatic potential energies between atoms and molecules, with mutual distances ranging from a few tens

to a few hundreds Bohr radii (Lepers & Dulieu, 2017). Representations should be able to account for long-range interactions.
8. Chemical and Physical Informedness: A representation is well-informed by chemistry or physics if it contains information about the chemical properties of each individual atom.

Polyatomic Complexes satisfy all these criteria (Khorana et al., 2024).

The software is helpful to researchers in cheminformatics, quantum chemistry, and materials science seeking a theoretically grounded approach to feature generation. Essentially Polyatomic Complexes are an alternative to the other common representations in cheminformatics which do not provide the same theoretical guarantees.

While widely used, current molecular representations—such as SMILES, SELFIES, molecular graphs, and ECFP fingerprints often fail to incorporate physically meaningful topological and electronic structure information (Khorana et al., 2024; Krenn et al., 2022; Manolopoulos & Fowler, 1992; Rogers & Hahn, 2010). These representations, although computationally efficient, are not designed to satisfy key scientific constraints such as topological accuracy, long-range interactions, and differentiability with respect to atomic coordinates (Bhadwal et al., 2023; Langer et al., 2022; Le et al., 2020).

Polyatomic Complexes address this gap by introducing a general-purpose, and topologically informed representation that integrates smoothly with modern ML pipelines. This fills a critical unmet need in the intersection of chemistry, materials science, and machine learning.

Moreover with the advent of topological deep learning such representations will become increasingly applicable (Zia et al., 2024). A classic example of this is Cellular Neural Networks (Khorana, 2024) or Cellular Gaussian Processes (Alain et al., 2024) which are compatible with Polyatomic Complexes (Khorana et al., 2024).

Code Contributions and Workflow

The provided code enables researchers to develop effective models for a variety of tasks in cheminformatics and materials science. The code/repository contributions can be summarized as follows:

1. An implementation of the core Polyatomic Complexes representation
2. An easy to use, well documented API for experiments. See the official documentation.
3. Easy integration with existing quantum chemistry libraries such as `pyscf` (Sun et al., 2020), and `pymatgen` (Ong et al., 2013).
4. Well packaged example datasets for baseline performance and benchmarking.

Moreover the Polyatomic Complexes are stratified into different categories depending on usage and experimental need. This is touched upon in the Software Description.

A standard workflow for molecular machine learning using Polyatomic Complexes would look as follows.

The Journal of Open Source Software

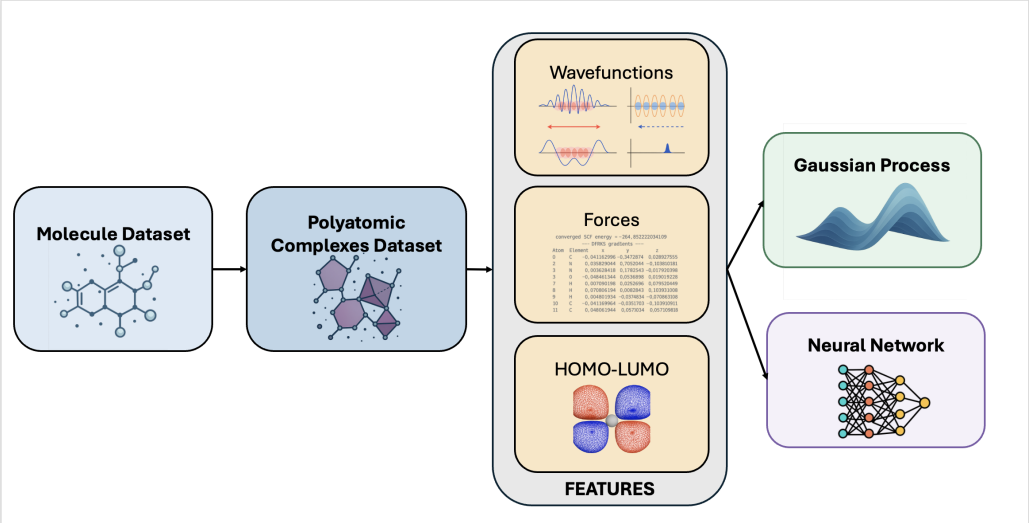

Figure 1: In this figure we see a standard molecular ML pipeline including Polyatomic Complexes.

The figure above shows the standard pipeline for many molecular machine-learning tasks. Initially, one receives a dataset consisting of both input and output columns. The input is usually a SMILES string (Weininger, 1988) or material, namely a PyMatgen `Structure` or `Molecule` (Jain et al., 2020). However, a wide variety of inputs are possible, such as molecular graphs, SEFLIES, and ECFP fingerprints (Krenn et al., 2022; Manolopoulos & Fowler, 1992; Rogers & Hahn, 2010). In the second stage, these input columns containing the molecule are transformed into Polyatomic Complexes. This enables one to compute numerous features ranging from purely topological or geometric features such as the Hodge Laplacians or spectral k-chains to force matrices and dipole moments. The third stage involves choosing a machine learning model and deciding which inputs to provide to it. Upon deciding on an architecture and features that suit the particular task, one trains their model and evaluates it.

Software Description

Our API is structured as follows:

1. `PolyatomicGeometrySMILE`: an interface for converting SMILES to polytomic complexes.
2. `AbstractComplex`: The base class and general purpose option.
3. `ForceComplex`: inherits from `AbstractComplex` and leverages methods from chemistry to provide detailed intermolecular force information.
4. `QuantumComplex`: inherits from `AbstractComplex` and leverages the B3LYP functional and DFT to provide highly accurate chemical information.
5. `QuantumWavesComplex`: inherits from `QuantumComplex` and provides long-range interactions and information about quantum wavefunctions.
6. `Datasets`: The general datasets API currently supports the ESOL, photoswitches, FreeSolv, and Lipophilicity datasets.

The software is modular, extensible, and written in Python. It supports input from standard molecular formats and returns representations and features suitable for use with ML libraries.

Repository and Installation

The source code for Polyatomic Complexes is hosted on GitHub. The software is open source under the MIT license and tested via continuous integration using GitHub Actions. Installation

instructions, API documentation, and tutorials are available at https://rahulkhorana.github.io/PolyatomicComplexes/

Benchmarks

Benchmark experiments and performance evaluations comparing Polyatomic Complexes to existing molecular representations (e.g., SMILES, SELFIES, ECFP) are provided in Khorana et al. (2024). These results include standard datasets in molecular property prediction and are fully reproducible via scripts and notebooks in the repository.

Acknowledgements

We would like to acknowledge Dr. Jin Qian, Dr. Marcus Noack, and others in the Chemical Sciences division at Lawrence Berkeley National Laboratory for their support, suggestions, and mentorship during the genesis of this project.

References

Alain, M., Takao, S., Paige, B., & Deisenroth, M. P. (2024). *Gaussian processes on cellular complexes*. https://doi.org/10.48550/arXiv.2311.01198

Bhadwal, A. S., Kumar, K., & Kumar, N. (2023). GenSMILES: An enhanced validity conscious representation for inverse design of molecules. *Knowledge-Based Systems*, *268*, 110429. https://doi.org/10.1016/j.knosys.2023.110429

Jain, A., Montoya, J., Dwaraknath, S., Zimmermann, N. E., Dagdelen, J., Horton, M., Huck, P., Winston, D., Cholia, S., Ong, S. P., & others. (2020). The materials project: Accelerating materials design through theory-driven data and tools. *Handbook of Materials Modeling: Methods: Theory and Modeling*, 1751–1784. https://doi.org/10.1007/978-3-319-42913-7_60-1

Khorana, R. (2024). *CW-CNN & CW-AN: Convolutional networks and attention networks for CW-complexes*. https://doi.org/10.48550/arXiv.2408.16686

Khorana, R., Noack, M., & Qian, J. (2024). *Polyatomic complexes: A topologically-informed learning representation for atomistic systems*. https://doi.org/10.48550/arXiv.2409.15600

Krenn, M., Ai, Q., Barthel, S., Carson, N., Frei, A., Frey, N. C., Friederich, P., Gaudin, T., Gayle, A. A., Jablonka, K. M., & others. (2022). SELFIES and the future of molecular string representations. *Patterns*, *3*(10). https://doi.org/10.1016/j.patter.2022.100588

Krenn, M., Häse, F., Nigam, A., Friederich, P., & Aspuru-Guzik, A. (2020). Self-referencing embedded strings (SELFIES): A 100. *Machine Learning: Science and Technology*, *1*(4), 045024. https://doi.org/10.1088/2632-2153/aba947

Langer, M. F., Goeßmann, A., & Rupp, M. (2022). Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning. *Npj Computational Materials*, *8*(1), 41. https://doi.org/10.1038/s41524-022-00721-x

Le, T., Winter, R., Noé, F., & Clevert, D.-A. (2020). Neuraldecipher–reverse-engineering extended-connectivity fingerprints (ECFPs) to their molecular structures. *Chemical Science*, *11*(38), 10378–10389. https://doi.org/10.1039/D0SC03115A

Lepers, M., & Dulieu, O. (2017). *Long-range interactions between ultracold atoms and molecules*. https://doi.org/10.48550/arXiv.1703.02833

Manolopoulos, D. E., & Fowler, P. W. (1992). Molecular graphs, point groups, and fullerenes. *The Journal of Chemical Physics*, *96*(10), 7603–7614. https://doi.org/10.1063/1.462413

Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V. L., Persson, K. A., & Ceder, G. (2013). Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, *68*, 314–319. https://doi.org/10.1016/j.commatsci.2012.10.028

Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, *50*(5), 742–754. https://doi.org/10.1021/ci100050t

Sun, Q., Zhang, X., Banerjee, S., Bao, P., Barbry, M., Blunt, N. S., Bogdanov, N. A., Booth, G. H., Chen, J., Cui, Z.-H., & others. (2020). Recent developments in the PySCF program package. *The Journal of Chemical Physics*, *153*(2). https://doi.org/10.1063/5.0006074

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, *28*(1), 31–36. https://doi.org/10.1021/ci00057a005

Zia, A., Khamis, A., Nichols, J., Tayab, U. B., Hayder, Z., Rolland, V., Stone, E., & Petersson, L. (2024). Topological deep learning: A review of an emerging paradigm. *Artificial Intelligence Review*, *57*(4), 77. https://doi.org/10.48550/arXiv.2302.03836